DNLP 作业 1

Zipf's Law 的中文语料库验证与中文的平均信息熵

朱道涵 20231202
20231202@buaa.edu.cn

Abstract

通过中文语料库来验证 Zipf's Law，并计算中文(分别以词和字为单位) 的平均信息熵。

Part1：Zipf's Law

Introduction

Zipf's Law（齐夫定律）是一种关于词频与词汇排序之间关系的经验定律。它由美国语言学家乔治·金斯利·齐普夫（George Zipf）在 20 世纪初提出。这个定律描述了在自然语言中，某个词的频率与其在频率排序表中的排名成反比关系。具体来说，Zipf's Law 表示为：如果将一段文本中的词按出现频率从高到低进行排序，那么第 i 个最常出现的词的出现频率将与第一个最常出现的词的频率成比例关系。

Methodology

首先将中文语料库中的文本进行合并，利用 jieba 库对合并后的文本进行分词，去掉停用词；随后统计词频并排序，绘制对数尺度的词频-次序函数曲线。

Experimental Studies

词频-次序函数曲线如图 1 所示。在对数尺度下，二者大致呈现负相关的正比例函数曲线，这充分说明 Zipf's Law 的正确性。

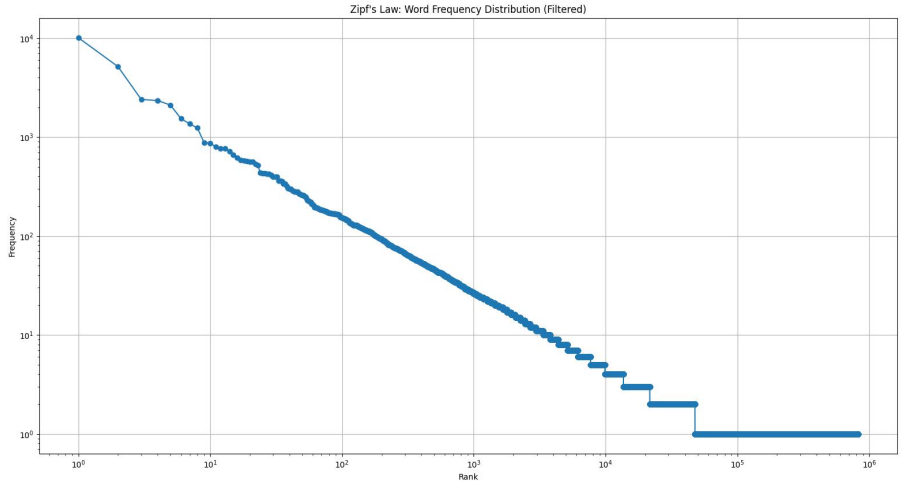

图 1：中文语料库的 Zipf's Law 曲线

Part2：信息熵

Introduction

信息熵用于表示信息的不确定性，熵值越大，则信息的不确定程度越大。数学公式为：

$$H(x) = \sum_{x \in X} P(x) log(\frac{1}{P(x)}) = -\sum_{x \in X} P(x) log(P(x))$$

规定：0 log（0） = 0。
联合信息熵可以用于估计二元模型，数学公式为：

$$\begin{aligned} H(X|Y) &= -\sum_{y \in Y} P(y) log(P(x|y)) \\ &= -\sum_{y \in Y} P(y) \sum_{x \in X} P(x) log(P(x|y)) \\ &= -\sum_{y \in Y} \sum_{x \in X} P(x) P(y) log(P(x|y)) \\ &= -\sum_{y \in Y} \sum_{x \in X} P(x, y) log(P(x|y)) \end{aligned}$$

三元模型的联合信息熵类似，不再赘述。
假设语言模型是一个 N-1 阶马尔科夫链，且可以利用 N-gram 模型进行建模，于是可以使用类似方法求出某种语言的信息熵。

Methodology

给定一个句子序列：

$$S = W_1, W_2, \ldots, W_K$$

对于一元模型，它的概率数学公式可以表示为:

$$P(S) = P(W_1, W_2, \ldots, W_K) = p(W_1)P(W_2|W_1)\ldots P(W_K|W_1, W_2, \ldots, W_{K-1})$$

对于二元模型，它的概率数学公式可以表示为:

$$P(S) = P(W_1)P(W_2|W_1)P(W_3|W_2)\ldots P(W_i|W_{i-1})\ldots P(W_n|W_{n-1})$$

对于三元模型，它的概率数学公式可以表示为:

$$P(S) = P(W_1)P(W_2|W_1)P(W_3|W_1, W_2)\ldots P(W_i|W_{i-2}W_{i-1})\ldots P(W_n|W_{n-2}, W_{n-1})$$

Experimental Studies

分别以词和字为单位的中文平均信息熵为单位如表 1 所示。以词为单位是通过 jieba 分词实现的，以字为单位则简单将字符串转化为列表即可。

表 1: 两种单位的中文语料库信息熵

单位 / 模型	词	字
1-gram	12.01262	9.50147
2-gram	6.89152	6.68503
3-gram	2.41693	3.94331

Conclusions

本次作业利用中文语料库，验证了中文语料库场景下 Zipf's Law 的正确性。估算了词与字两种单位的中文语料库信息熵。